Automatic standardization of questions in Frequently Asked Questions

Hideki Sugizuru

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 1999

Keywords: FAQ, standard form, edit, information extraction, summary.

Recently, on the Internet, there are many FAQs, a FAQ is a collection of typical questions and their answers in certain field. Usually, such FAQs are complied by hands. Because compiling and updating FAQs by hand is time-consuming, we have been studying the method of making collections questions and answers automatically from Netnews newsgroup fj.sys.sun; the output of this study is presented as Sun QA-Pack on WWW.

Sun QA-Pack uses the important sentences which are extracted from a question article of fj.sys.sun as a headline of the article. That is, expression of summary sentences is not changed from the original sentence; the quality of summary sentences are dependent on the person who wrote the article.

In this research, I studied the method of standardize the summary sentences of Sun QA-Pack. If standardized, two sentences the express the same question should be the same. Consistency can be given to Sun QA-Pack if all summary sentences are changed into standard forms.

First , I studied the features of summary sentences and classified. The summary sentences was into four types by contents: "したい (want)"," できない (can't)"," 教えて下さい (teach me)"," 状況説明 (situation-explanation)". Based on this study, I set two standard forms as the following :

1. (noun-phrase) *wo* (verb) *shi-tai*
   (I want to (verb)(noun-phrase))

2. (noun-phrase) *ga* (verb) *deki-nai*
   (I can't (verb)(noun-phrase))

The summary sentences of the "したい"(want) type are standardized into the first form. The summary sentences of the "できない"(can't) type are standardized into the second form. The summary sentences of the "教えて下さい"(teach me) type are the cases that we standardize into the first form by inserting the verb " 知る"(know); thus, these sentences become "I want to know (noun-phrase)". The summary sentences of the "状況説明"(situation-explanation) type is kept out from this research because since there is no clear feature.

Second, I implemented the standardization system of the question sentences. This system consists of three modules: the adjustment module, the standardization module, and the output-selection module.

The adjustment module eliminates tags of domain-specific terms in the summary sentence, divides the sentences, adjust it, and hands to the standardization module. The domain-specific term tags are used for automatic classification of Sun QA-Pack; this system eliminates them because they are not necessary. Since two or more sentences are often used as the summary, this module divides sentences into a sentence using periods.

The standardization module first judges the type of the sentence: then, it standardize the sentence. Sentence types are judged using expression patterns in the sentence end. Sentences are standardized as the following:

1. The module decides the standardization rules to apply. The standardization rules are made using the relation between the distinctive expression patterns of question sentences and the standard forms of question sentences. This module uses eight standardization rules.

2. The module uses Japanese morphological analysis system **Juman** to extract verbs and objects.

3. The module extracts important verbs using the standardization rules and the expressions of "intension" and "negation", and related objects to the verb using standardization rules and particles.

4. The modules standardize the sentence by applying the standardized rule to above verbs and objects.

The output-selection module selects the most appropriate sentence if the standardization module outputs two or more standardized sentences.

I evaluated the system using 215 summaries of question articles in the Netnews newsgroup fj.sys.sun. The situation-explanation type summary sentences are omitted previously because this type is not the target of this research; obscure summary sentences are also omitted previously.

The standardization module standardized 124 summaries (47 percent) correctly, 53 summaries (20 percent) incorrectly. The module failed to standardize 87 summaries (33 percent). Avoiding the summaries that were not standardized, the module achieved 70 percent of accuracy.